On the mechanisms of protein interactions

Marín-López, Manuel Alejandro; Planas-Iglesias, Joan; Aguiree-Plans, Joaquim; Bonet, Jaume; Garcia-Garcia, Javier; Fernandez-Fuentes, .Narcis; Oliva, Baldo

Published in:
Bioinformatics

DOI:
10.1093/bioinformatics/btx616

Publication date:
2017

Citation for published version (APA):
Marín-López, M. A., Planas-Iglesias, J., Aguiree-Plans, J., Bonet, J., Garcia-Garcia, J., Fernandez-Fuentes, . N., & Oliva, B. (2017). On the mechanisms of protein interactions: predicting their affinity from unbound tertiary structures. *Bioinformatics*, *34*(4), 592-598. https://doi.org/10.1093/bioinformatics/btx616

Bioinformatics, YYYY, 0–0
doi: 10.1093/bioinformatics/xxxxx
Advance Access Publication Date: DD Month YYYY
Manuscript Category

Structural Bioinformatics

On the mechanisms of protein interactions: predicting their affinity from unbound tertiary structures

Manuel Alejandro Marín-López[1], Joan Planas-Iglesias[2,*], Joaquim Aguirre-Plans[1], Jaume Bonet[3], Javier Garcia-Garcia[1], Narcis Fernandez-Fuentes[4] and Baldo Oliva[1]*

[1]Structural Bioinformatics Lab, Department of Experimental and Health Science, Universitat Pompeu Fabra, Barcelona 08005, Catalonia, Spain.

[2]Division of Metabolic and Vascular Health, University of Warwick, Coventry CV4 7AL, UK.

[3]Laboratory of Protein Design & Immunoenginneering, School of Engineering, Ecole Polytechnique Federale de Lausanne, Lausanne 1015, Vaud, Switzerland.

[4]Institute of Biological, Environmental and Rural Sciences, Aberystwyth University, Aberystwyth SY23 3DA, UK.

*Both authors wish to be considered co-corresponding.

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: The characterization of the protein-protein association mechanisms is crucial to understanding how biological processes occur. It has been previously shown that the early formation of non-specific encounters enhances the realization of the stereospecific (i.e. native) complex by reducing the dimensionality of the search process. The association rate for the formation of such complex plays a crucial role in the cell biology and depends on how the partners diffuse to be close to each other. Predicting the binding free energy of proteins provides new opportunities to modulate and control protein-protein interactions. However, existing methods require the 3D structure of the complex to predict its affinity, severely limiting their application to interactions with known structures.

Results: We present a new approach that relies on the unbound protein structures and protein docking to predict protein-protein binding affinities. Through the study of the docking space (i.e. decoys), the method predicts the binding affinity of the query proteins when the actual structure of the complex itself is unknown. We tested our approach on a set of globular and soluble proteins of the newest affinity benchmark, obtaining accuracy values comparable to other state-of-art methods: a 0.4 correlation coefficient between the experimental and predicted values of ΔG and an error < 3 Kcal/mol.

Availability: The binding affinity predictor is implemented and available at http://sbi.upf.edu/BADock and https://github.com/badocksbi/BADock

Contact: baldo.oliva@upf.edu; j.planas-iglesias@warwick.ac.uk

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Proteins are the building blocks needed by living organisms to carry out most of their cellular processes. To fulfill their functional role, proteins need to physically interact with one another (as well as with other biomolecules, e.g. DNA) forming transient and permanent complexes in a time and location dependent manner (Gavin, et al., 2002; Robinson, et al., 2007). Hence, the characterization of binding affinities and molecular mechanisms of protein-protein associations are critical challenges in current biomedical research. The formation of a protein complex involves three steps: the initial fast formation of a non-specific encounter complex from free proteins in solution, the two-dimensional search on both protein surfaces that brings the pair to an orientation close to the native complex (transient complex), and the subsequent conformational changes (Schreiber, et al., 2009). The conformations of the dynamic contacts of proteins with other proteins (encounter complexes) can be stabilized through long-range electrostatic interactions and possibly supplemented by short-range interactions(Tang, et al., 2006). In fact, studies in protein design show that association rates can be increased by optimizing the electrostatic attraction between proteins (Selzer, et al., 2000). Thus, the ensemble of encounter complexes is crucial to accelerate the association process. Conversely, large conformational changes upon binding slow down the association process (Zhou and Bates, 2013). This model of the protein binding mechanism helps to contextualize different published results on predicting protein-protein interactions that involve non-interacting regions. For example, we developed a protein interaction predictor relying on the classification of structural domains (Andreeva, et al., 2008) and super-secondary structures (Bonet, et al., 2014) where a relevant number of such structural features were located outside the binding interface (Planas-Iglesias, et al., 2013; Planas-Iglesias, et al., 2013). In a different context, when only the structure of the two unbound proteins that form a binary complex is known, a docking strategy to predict the complex structure can be used, producing several candidates that are ranked according to a certain scoring function (Feliu, et al., 2011; Feliu and Oliva, 2010; Segura, et al., 2015). Indeed, protein docking and the development of scoring functions to rank docking models is a fertile ground as proved by the extensive literature of proposed methods and an active CAPRI and CASP-CAPRI competition (see reviews by Lensink et al. (Lensink, et al., 2016), Gromiha et al. (Gromiha, et al., 2016) and references therein). Wass et al. showed that sets of docking poses could be used to discern between interacting and non-interacting proteins through the presence of near-native decoys and the distribution of docking scores (Wass, et al., 2011). All these findings support the funnel-like intermolecular energy landscape theory for molecular interactions (McCammon, 1998), and hint at the existence of a common feature or profile for interacting proteins, as if their recognition is not only dependent on the specific binding interface.

The energy landscape of protein interactions is also characterized by their association rate, which along with the dissociation rate depicts the binding affinity of the protein complexes. Such affinity is described by the equilibrium dissociation constant (Kd), and from a thermodynamic perspective —assuming the standard concentration of 1mol/dm^3 and equaling the quotient of activity factors to 1, is calculated by the Gibbs free energy using the formulae: $\Delta G = -RT\log_e(Kd)$. Experimental techniques for measuring binding affinity are expensive and time-consuming (Garcia-Garcia, et al., 2012). For this reason, many computational methods have been developed in the last decades to predict the binding affinity (Horton and Lewis, 1992; Kastritis, et al., 2014; Ma, et al., 2002; Moal, et al., 2011; Vangone and Bonvin, 2015), and only few have considered the effect of non-binding regions (Tian, et al., 2012). However, most of these methods show poor accuracy when tested against large datasets (Kastritis and Bonvin, 2010). Such methods usually rely on the known structure of binary complexes (Erijman, et al., 2014), and have eventually proved the relevance of the quality of the crystal structure of the complex to improve the prediction (Marillet, et al., 2016). The affinity prediction for the complex is achieved by identifying features on the native interface and applying scoring functions (Moal, et al., 2011), either based on statistical potentials (Su, et al., 2009), on atomic physical interactions (Audie and Scarlata, 2007) or complementarities in the surfaces obtained by docking approaches (Vreven, et al., 2012). To account for conformational changes, often linked to protein interactions, molecular dynamics simulations and simplified models, such molecular mechanics Poisson-Boltzmann surface area and Generalized Born variant, provides a valid, albeit more computationally expensive, route to improve the prediction of the binding energy between proteins (Gohlke, et al., 2003; Gumbart, et al., 2013; Moritsugu, et al., 2014; Rodriguez, et al., 2015).

Questions on the role of non-interacting regions affecting the binding affinity and the energy landscape of protein-protein interactions have been addressed only of late (Kastritis, et al., 2014; Tian, et al., 2012). Still, even these recent methods use the structure of the protein complex to calculate the long-distance interaction between the residues of both partner proteins and their opposite native interfaces (Kastritis, et al., 2014; Vangone and Bonvin, 2015) and hence have limited applicability. With the aim to shed light into the role of non-interacting sites, we study the formation and binding affinity of binary complexes of globular soluble proteins. We use the poses resulting from the protein-protein docking search to scout the conformational space of potential encounter complexes. We classify the docking space into different types of productive and non-productive conformations according to their potential to form the native structure of the binary complex. Based on this analysis we endeavor to predict the binding affinity using the unbound protein structures, proving its feasibility. We have tested our approach using the affinity benchmark 2 (Vreven, et al., 2015), the largest affinity benchmark up to date. In contrast to current state-of-the-art methods that require the native structure of the binary complex, we conclude that only the structure of the unbound partners is required, thus extending the applicability of predictions despite lowering the quality on the prediction but with a reasonable margin of error (in most cases lower than 3Kcal/mol).

2 Methods

2.1 Datasets

We use the Docking Benchmark 5 and the Binding Affinity Benchmark 2 (Vreven, et al., 2015) to study the conformational space of docking poses resulting from docking experiments. The benchmarks respectively consist of 230 and 179 non-redundant high quality structures of protein

complexes classified by biological functions. The sets are divided in three categories: enzymes, antibody-antigen and others (including membrane-bound receptors, G-protein (or G-protein-coupled receptor) proteins and a set of miscellaneous protein types and functions). In addition, for each protein the interface-RMSD (Méndez, et al., 2003) is reported. This measure can be used to estimate the degree of conformational change that a protein undergoes upon binding, allowing to split the datasets into rigid (interface-RMSD<1Å) and flexible (interface-RMSD≥1Å) interactions. We restrict our dataset to globular soluble proteins by omitting the categories of membrane-binding receptors and G-proteins (or G-protein-coupled receptors). We also omitted antibody-antigen complexes as we considered these to be a particular case of protein-protein interactions which mechanisms of recognition and binding may be more intriguing (see supplementary material). The trimmed datasets are referred here as DB5 (from Docking Benchmark 5) and AB2 (from Affinity Benchmark 2). The analyses with different scoring functions were performed on 94 complexes out of the AB2 dataset that are also found in the CCHarPPI server (Moal, et al., 2015).

2.2 Docking, refinement and scoring

PatchDock (Schneidman-Duhovny, et al., 2005) was used with default parameters to obtain the docking poses (or decoys), which were ranked according to a geometric shape complementary score (rigid docking). Poses were obtained by docking the conformations of the two interacting proteins in its bound form (i.e. uncoupling the complex and trying to reconstitute it by docking). However, decoys were refined and rescored using FiberDock (Mashiach, et al., 2010; Mashiach, et al., 2010) to simulate the flexibility, optimize the interaction and calculate the affinity of the interaction with all decoys, near-native and non-native poses, under the same conditions. Besides, for the analysis of the prediction of binding affinity, we performed the docking on the unbound forms of the complex. Finally, all docking poses were scored by three statistical potentials: EPAIR, ES3DC and E3D from Feliu et al. (Feliu, et al., 2011). EPAIR is the classical statistical potential for the interaction of two residues. ES3DC is a refinement of EPAIR that considers the condition in which the residues sit (secondary-structure and degree of accessibility). The last scoring term, E3D, concerns only the distance at which pairs of residues interact and increases together with the number of interacting residue-pairs, thus reflecting the size of the interface.

2.3 Classification of docking poses in encounter complexes

First, we assume that PatchDock samples sufficiently the conformational space of encounter complexes. Then, we classify the obtained poses into four different classes: Near-Native, Face-Face, Face-Back and Back-Back, reflecting the relative positions of the binding sites of each protein partner. Near-Native class correspond to all docking poses with a ligand-RMSD < 10A, ligand-RMSD being the RMSD of the ligand coordinates after superposition of the receptor coordinates (Méndez, et al., 2003). When the ligand-RMSD is larger, the classification depends upon the accessibility of the interacting interfaces of the partners: Face-Face are docking poses in which the binding sites of both protein-partners face each other (i.e. they are inaccessible to other proteins); Face-Back, when only one binding site interacts with the protein partner (i.e. the binding site of one of the proteins is freely accessible); and Back-Back, when both binding sites are free to interact with other protein units (see example in supplementary figure S1). To elucidate if a binding site of one of the protein partners (A) remains accessible in a complex decoy or pose (formed by A and its partner B), a guided docking using PatchDock is done between the pose and the single chain of the other protein partner (B). The docking is guided using the native interface residue-residue distance constraints between proteins in the decoy; all other parameters were set as per default. If PatchDock guided docking produces results, the binding site of the tested partner (A) is still accessible in the docking pose; otherwise, the binding site is not accessible and thus the protein partner B in the decoy is placed totally or partially on top of the binding site. This procedure is done twice, once for each protein partner (A and B) to determine the accessibility of both binding sites and to classify each docking pose in one of the non Near-Native aforementioned classes. If the binding site of both partners, A and B, is not accessible, then the docking poses need small rotations to produce a near native solution, we classify such poses as Face-Face. If both binding sites are accessible in the decoy, then the orientation is opposite to the native orientation and we classify it as Back-Back. Otherwise, if one of the partners, A or B, has the interface inaccessible and the pose is classified as Face-Back.

2.4 Correlations and predictions

We use the absolute values of Pearson's correlation coefficients (R) to determine the linear dependence between the scores of different classes of docking poses considering only one score at a time or multiple scores. The score of a class or group of conformations is obtained by averaging all the poses in the group. We use linear regression models for predicting the affinity (ΔG) from the unsolved forms of the interactions in the AB2 dataset. The models are trained and tested using Scikit-learn module of python (Pedregosa, et al., 2011) with the docking scores. We randomly split the data into 10 subsets to perform a 10-fold cross-validation analysis; the procedure was repeated 1000 times. We also use the Pearson's correlations coefficients between the experimental and predicted affinities, and prediction ratio (as defined in (Marillet, et al., 2016), see also supplementary table S5) to evaluate predictions. The docking scores of the predictions were obtained with the docking poses of the bound and unbound conformations for testing the differences. Any of the Pearson's correlations is assumed to be statistically significant if the associated p-value is lesser than 0.05. The significance of differences in the performance of the models applied to native complexes or to docking decoys was assessed using a Mann-Whitney test and the Hodge-Lehman estimate of the population shift. Density plots and distributions of scores are obtained with Seaborn and Matplotlib modules of Python.

3 Results

3.1 Analysis of the conformational space of encounter complexes

To decipher the potential role of docking poses in the encounter of two proteins, we followed a strategy consisting of: a) uncouple each complex in the DB5 dataset, b) rebuild complexes using PatchDock, c) refine results with FiberDock, d) score the complete set of solutions using different energetic terms and statistical potentials (including EPAIR, ES3DC and E3D) and e) classify poses in four classes: Near-Native (NN), Face-Face (FF), Face-Back (FB) and Back-Back (BB), depending on the relative position of the binding sites (see Methods). Our starting assumption is that defined classes (step e) represent four conformational macro-states of the interaction: the first two correspond to productive encounters of the interacting partners and the last two to the non-productive ones. To test this hypothesis, we initially analyzed the span of scores (step d) within each class and compared the different classes using the arithmetic mean. Figure 1 summarizes this analysis for ES3DC score.

Since docking scores are designed to rank near-native poses, it is not surprising that Near-Native scores are much smaller than the other groups. Distributions of E3D, EPair and FiberDock scores show a similar trend (Supplementary Figures S2, S3 and S4, respectively). Interestingly, these scores describe a decreasing slope from non-productive to productive conformations, where the differences between the BB and the FF groups are much smaller than those existing between the FF and the NN classes. Considering that all docking solutions were included in the analysis regardless of their conformity with the crystallographic dimer (supplementary Figure S5), this trend supports the previously reported funnel like model for molecular interactions (Planas-Iglesias, et al., 2013; Wass, et al., 2011), which proposes that the interacting partners explore a wide conformational and (high-) energetic space before committing themselves into the interaction.

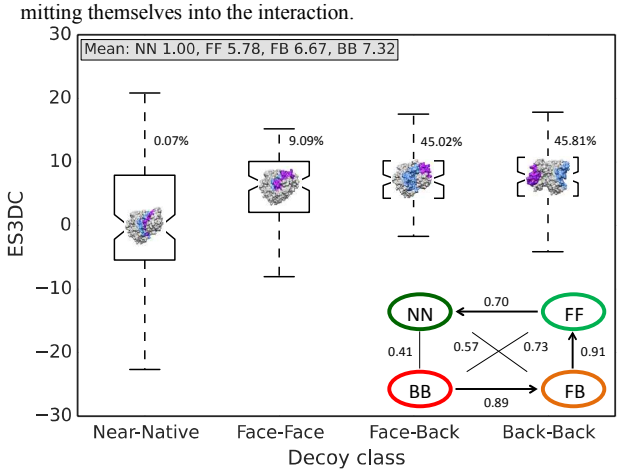
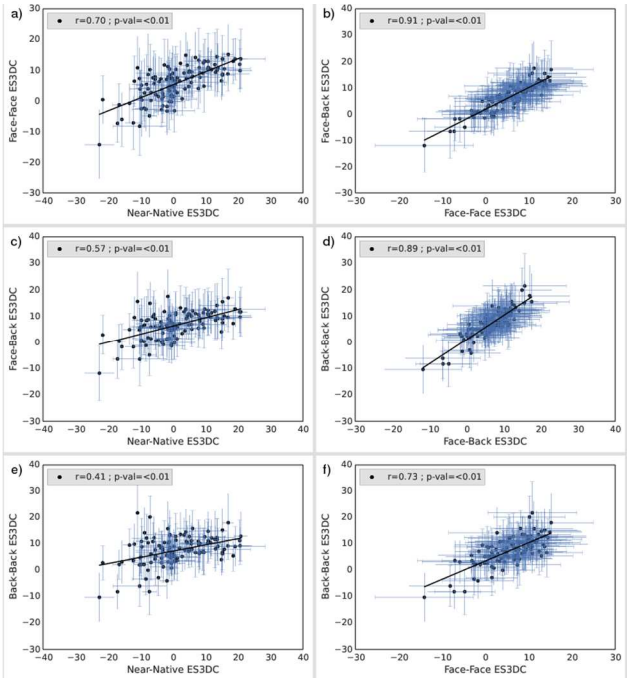


Figure 1. Boxplots of ES3DC averaged scores of several protein-protein interactions. Boxplots represent the distributions of the average of ES3DC scores in the NN, FF, FB and BB classes for the protein interactions of the DB5 dataset. Values next to each box show percentage of decoys of each class. Mean values for each class are shown in the gray legend at the top. A representative decoy is shown inside each boxplot (see Supplementary Figure S1). The inner plot in the bottom-right shows a directed graph inferring the binding process directionality, based on the correlations (see legend of Figure 2).

We have shown a funnel-like trend for a population of different pairs of proteins, but whether the same holds true for individual pairs needs to be tested. Hence, for each protein pair we calculated the Pearson's correlation between the average of the docking scores of each class. Figure 2 shows the results for ES3DC score (other scores in Supplementary Figures S6-S8), and the fitted values of all regression models are shown in Supplementary Table S1. BB energy-scores are often higher than FB ones, these higher than FF scores, and the NN class always contains the lowest of them. Two observations are noteworthy. First, FF, BB and FB groups of poses strongly correlate between them; conversely, the NN group of poses has a weaker correlation with the rest. These correlations are weaker for FiberDock scores, but all correlations are significant (p-values lower than 0.01) regardless of the scoring approach. Second, the span of scores within each decoy class is very large because of the high variability of surfaces in different complexes. Therefore, some scores in the BB or FB orientations (i.e. non-productive encounters) can even have lower energy than a Near-Native solution. Error bars in Figure 2 (and supplementary figures S6-S8) show the magnitude of this variability, despite of which a significant correlation between different decoy classes is still preserved. This is expected for NN and FF poses, since the relatively correct orientation of the encounter should be effective even if it does not result in an immediate complex. In such cases, the scouting of the conformational space is restricted to small rigid rotations while the partners do not need to be physically separated, reducing the time required to find the native form of the complex. The observed correlation between non-productive and productive orientations is however unexpected. However, since the scoring trend is preserved (high or low) regardless of the docking pose productivity class, non-productive orientations should describe the affinity of the molecular association as well as the productive ones.

Hence, a logical route can be traced from the BB poses to the NN orientations, where we consider that each class is a macro-state of the binding process. Such a route can be described as a graph in which each macro-state is represented as a node and the transitions between them as edges. We apply data processing inequalities (Margolin, et al., 2006) to reconstruct the network that connects the groups of poses using correlations between energies instead of mutual information: correlations between the energies of directly connected nodes must be higher than between nodes connected indirectly (i.e. by a transitive relationship). Pearson's correlation between the scores of these classes support a model that correlatively connects BB, FB, FF and NN classes (see inward graph in Fig. 1 for ES3DC scores; ibid. in Supplementary Figs. S2-4 for E3D, EPAIR and FiberDock respectively). The same trend is observed if correlation slopes



(Supplementary Table S1) are considered. Therefore, from our results we infer a path connecting the non-productive and the productive states, where Face-Back and Face-Face play a potential mechanistic role drawing near the binding sites of the two interacting partners. This model concurs with a very recent modelling experiment deciphering the association dynamics of the bacterial ribonuclease barnase with its inhibitor barstar (Plattner, et al., 2017). Plattner et al. show that initial steps towards binding also involve conformations that we defined as BB and FB.

Figure 2. Scatterplot of ES3DC averaged scores between decoy classes. Each dot shows the relationship between the averages of the ES3DC scores of poses with different decoy conformational classes (standard deviations are shown in error bars): NN vs. FF (a); FF vs. FB (b); NN vs. FB (c); FB vs. BB (d); NN vs. BB (e); FF vs. BB (f). Pearson's correlations are shown in the legends at the top of each scatterplot (they are used in the directed-graph in figure 1). Least squares fitting curve is shown (slope and y-coordinate interception are in supplementary table S1 for the sake of comparison).

3.2 Docking scores correlate with binding affinities on all different classes of docking poses

Current approaches to predict the binding affinity between two proteins rely on several scores and energies computed on the 3D structure of the native conformation of its binary complex (Moal, et al., 2011; Moal and Bates, 2012). Our previous analyses suggest that, if the scores of the Near-Native can be used to predict the affinity, then the scores of the rest of the classes might be used as well. To prove this hypothesis, we calculate on the unbound pairs in DB5 and AB2 ($DB5 \cap AB2$) the Pearson's correlation between the average of the scores of each class and the experimentally determined ΔG (see Table S2). FiberDock scores of the Near-Native poses significantly correlate ($p < 0.05$) with the affinity (ΔG) as previously reported (Vangone and Bonvin, 2015). Interestingly, FiberDock scores of the non-productive orientations significantly ($p < 0.05$) correlate with ΔG too. More importantly, the electrostatic terms of FiberDock significantly correlate with the affinity ($p < 0.05$) in all classes except for the NN, where van der Waals and de-solvation energies have a major role. These results agree with earlier studies suggesting that electrostatic forces dictate the formation of encounter complexes (Alsallaq and Zhou, 2008; Zhou and Bates, 2013; Zhou, 1993). This is a trend preserved in interactions between proteins and other biomolecules such as nucleic acids (Fornes, et al., 2014) and lipids (Barneda, et al., 2015; Planas-Iglesias, et al., 2015), suggesting that the role of non-interacting regions of proteins in such intermolecular interactions could also be relevant. Notably, we don't need to simulate the dynamics of the protein-protein encounters to reach a similar conclusion; instead we rely only on a limited exploration of the conformational space of the interaction represented by several docking solutions.

The statistical potentials EPAIR and ES3DC indicate the active role of specific residue-pairs in the interface of known interactions, while E3D is directly proportional to the number of residues in contact (i.e. the interface size). Congruently, correlations between ΔG and the average of E3D scores are higher for the productive (NN and FF) than the non-productive (FB and BB) orientations. There is also a high and significant correlation for all orientations of both EPAIR and ES3DC with the affinity, showing that specific residues from any location on the protein surface may have a role in binding.

All these results suggest that proteins increase their affinity by both lowering the energy of the stereospecific native complex and enhancing the encounter complexes in any of its potential different orientations.

3.3 Binding affinity can be predicted using all docking poses

We have shown in the previous analysis how ΔG correlates with the average values of the statistical potentials EPAIR, ES3DC and the electrostatic terms of the FiberDock score for all groups of decoy classes. Hence, we test the Pearson's correlation between the ΔG and the averages computed on all the decoys of a docking (see table S2). The scores of FiberDock, E3D and ES3DC show a significant correlation with ΔG when using the native complex structures of the AB2 dataset, as expected from previous works (Vangone and Bonvin, 2015). In comparison, when using all the available poses resulting from unbound docking, the average of the electrostatic terms of FiberDock and the averages of the statistical potentials EPAIR and ES3DC significantly correlate with ΔG ($p < 0.05$). Furthermore, when the AB2 dataset is split into rigid and flexible cases (see Methods), the average of FiberDock scores (obtained with all decoys of each protein-pair) is significantly correlated ($p < 0.05$) to ΔG only in rigid cases. Interestingly, the average of the statistical potentials EPAIR and ES3DC are correlated with ΔG in both cases, rigid and flexible, with most points within a margin error of 2.8 Kcal/mol (Table S2, Figure S9). According to previously reported results (Horton and Lewis, 1992; Zhou and Bates, 2013), the correlation of the binding affinity with van der Waals terms shows the role of the surface complementarity, the solvation and the loss of entropy produced by the conformational accommodation of the protein-partners.

Complementarily, the role of the electrostatic potential terms should be more relevant for the recognition of the protein-partners (Schlosshauer and Baker, 2004). Hence, we have analyzed the correlation of the different types of scores with the k_{on} and k_{off}, similarly as for ΔG, for the number of cases in the AB2 dataset that these rate constants are determined. Both constants describe the protein-protein association rate kinetically, taking into account the diffusion-limited approach of the two interacting proteins and the stability of the intermediate interaction (Schreiber, et al., 2009). We hypothesize that the averages of the scores of the docking poses should better correlate with k_{on}, whereas using only the Native conformation to calculate the scores should correlate with k_{off} (Ubbink, 2009). However, the small size of the sample has limited our conclusions (Supplementary Table S3).

	AB2		AB2 Rigid		AB2 Flexible	
	Native	Poses	Native	Poses	Native	Poses
Fiber	0.30	0.29	0.41	0.37	0.21	0.23
VdW	0.37	0.14	0.50	0.20	0.32	0.06
Elec	0.16	0.34	0.17	0.43	0.12	0.24
HB	0.22	0.13	0.26	0.14	0.18	0.05
EPAIR	0.08	0.31	0.18	0.29	0.06	0.30
ES3DC	0.21	0.36	0.28	0.40	0.12	0.27
E3D	0.37	0.04	0.51	0.09	0.27	0.05

Table 1 Pearson's correlation between experimental and predicted ΔG. Scores of the native conformation of the complexes (Native) and the averages with all poses from a docking search with PatchDock (Poses) are shown. The first columns show the average of the correlations for the AB2 dataset. See error intervals (RMSE) in Table S4. The last groups of columns show the results split for rigid and flexible cases of the AB2 dataset. Correlations are shown for statistical potentials EPAIR, ES3DC and E3D, and for FiberDock scores (Fiber) also decomposed in van der Waals attractive (VdW), electrostatics attractive terms at long range (Elec) and hydrogen bonding energy terms (HB).

From these analyses, we infer that the average of scores of many different docking potentials, obtained with all the poses of a docking search between two proteins, can be used to predict their ΔG of binding form their tertiary structure (unbound forms). Specifically, one of the strongest correlations, obtained with ES3DC, is also very robust as it remains reliable for both flexible and rigid cases. We have created a linear regression model to predict ΔG using the ES3DC scores of all docking poses and a ten-fold validation protocol (see Methods, Supplementary Figure S10). We have applied a similar model to predict ΔG using only the native conformation. We have also generated similar models with other scores: EPAIR, E3D, hydrogen-bond, van der Waals and electrostatic terms of FiberDock (see Table 1 and supplementary Table S5). The addition of more terms to the linear models didn't improve the results, while unnecessarily increasing the overfitting of the model. Hence, we proceeded analyzing models which considered only one score. We have compared the results of using only the native conformation, where the best potentials are E3D, hydrogen-bond and attractive van der Waals terms of FiberDock, or all docking decoys, where the best potentials are ES3DC, EPAIR, and electrostatic terms of FiberDock. Interestingly, the

differences between the predicted and the measured binding affinities are comparable regardless of using only the native structure or the whole set of docking poses (Supplementary Table S6). We have also compared other potentials and scores from the results of the CCHarPPI server for 94 complexes of the AB2 dataset, although this approach could only be applied on the native conformation (Supplementary Tables S4 and S5). Finally, we further compared the use of unbound or bound conformations for the prediction, showing that both yielded comparable differences between the predicted and the measured binding affinities (Supplementary Figure S11, p=0.404 and p=0.391 for rigid and flexible cases, respectively).

The linear regression model obtained with the whole set of decoys from a docking search and the ES3DC statistical potential has a slope of 0.23 and intercepts at -12.3, and was obtained using all the data available in AB2. The predicted values of ΔG in the ten-fold cross-validation significantly correlate (0.36 average Pearson's correlation, $p < 0.05$) with the experimental, with an average error (RMSE) of 2.84 kcal/mol. Furthermore, two thirds of the assessed pairs obtain predictions within this range (38.30% of the pairs have predictions differing at most 1.4 Kcal/mol from experimental values). This effect is more noticeable in the flexible cases, where predictions are within 1.4 or 2.8 Kcal/mol for 43.48% and 76.09% of the cases, respectively (Supplementary Figure S6)

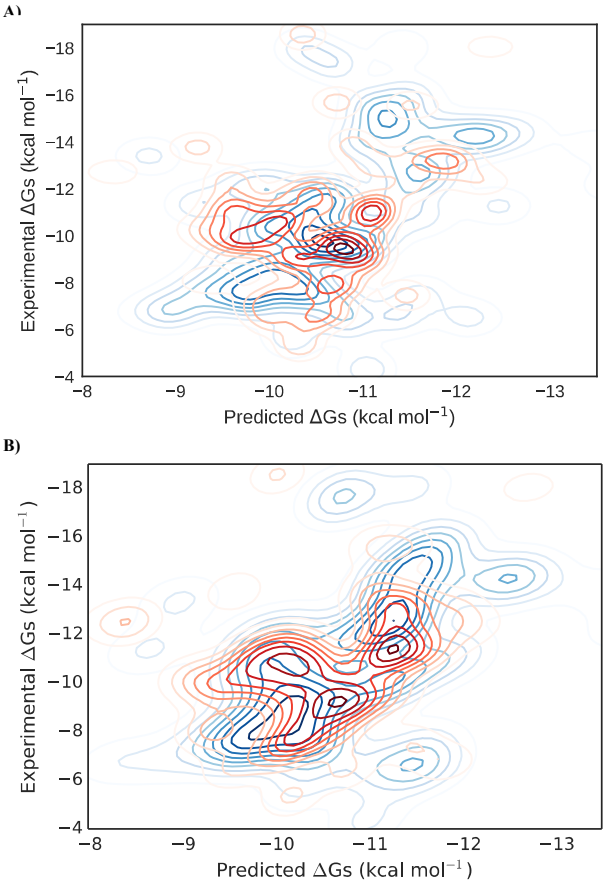


Figure 3. Density plot between experimental and predicted ΔG using the ES3DC statistical potential and all docking poses. Predictions are made using the test sets of 1000 random ten-fold cross validation models with the ES3DC averaged scores of all docking poses in the AB2 dataset using bound (A) or unbound (B) conformations. Blue and red lines show the density plot for rigid and flexible cases of AB2 respectivelly.

Figure 3 shows the density plot between predicted and experimental ΔG using the test sets of 1000 ten-fold cross-validation regression models, using both bound (A) and unbound (B) conformations of the proteins interaction. Differences between both conformations are quasi-negligible for flexible docking cases, proving the coherence of our approach to use the unbound conformations (see supplementary material). In comparison with other state-of-art approaches our method is less accurate (i.e. the Pearson's correlation between experimental and predicted ΔG of the most recent approaches ranges from 0.48 (Kastritis, et al., 2014) to 0.73 (Vangone and Bonvin, 2015). Other models for different types of proteins obtained correlations from 0.51 to 0.64 (Moal, et al., 2011). However, these approaches can only be applied if the structure of the binary complex is known, while ours only requires the structure of the unbound proteins or just the structure of the protein fragments or domains involved in the interaction. We provide the web-server Binding Affinity Dock (http://sbi.upf.edu/BADock), which implements the above described model for the prediction of binding energies of protein pairs. We also provide a github repository with the data and scripts to reproduce the work (https://github.com/badocksbi/BADock).

4 Conclusions

We have used the protein-docking method PatchDock to sample the conformational space of the non-specific complexes formed during the association process of two soluble and globular proteins. We have classified the decoys into four classes, depending on the orientation of the binding sites of the protein partners: two productive and two non-productive. We have shown that there is an association between the energetic terms and docking scores in all classes of conformations. A mechanistic path can be inferred from the direct-graph analysis of the correlations of the averages of docking scores. We have observed correlations between the experimental ΔG and the average of statistical potentials and electrostatic energy terms of the poses obtained by docking. The implication of electrostatic energies in the non-productive conformations agrees with previous studies that suggested that encounter (non-native) complexes are stabilized by these forces (Schlosshauer and Baker, 2004). Finally, we have developed a binding affinity predictor based on the whole set of docking poses, without requiring the structure of the complex. Although our method is less accurate than others it is still competitive, as it can cover many other proteins for which the structure of the complex is unknown, while achieving a relevant correlation between the prediction and the experimental value of ΔG. Nevertheless, we wish to note that when the native structure of the complex is known, many other approaches will obtain better accuracy than us.

Acknowledgements

MML acknowledges the fellowship from Generalitat de Catalunya (FI-DGR2012). Authors wish to acknowledge Dr. Ozlem Keskin and Dr. Attila Gursoy from KOC University for helpful discussions and Mr. James Mitchel from University of Warwick for proofreading the manuscript.

Funding

The work has been supported by grants BIO2014-57518-R and BIO2011-22568 of the Spanish Ministry of Economy (MINECO), INB 2015-2017 of ISCIII, and 2014SGR1161 of Generalitat de Catalunya

Conflict of Interest: none declared.

References

Alsallaq, R. and Zhou, H.-X. Electrostatic Rate Enhancement and Transient Complex of Protein-Protein Association. *Proteins* 2008;71:320-335.

Andreeva, A.*, et al.* Data growth and its impact on the SCOP database: new developments. *Nucleic Acids Res* 2008;36(Database issue):D419-425.

Audie, J. and Scarlata, S. A novel empirical free energy function that explains and predicts protein-protein binding affinities. *Biophysical Chemistry* 2007;129:198-211.

Barneda, D.*, et al.* The brown adipocyte protein CIDEA promotes lipid droplet fusion via a phosphatidic acid-binding amphipathic helix. *Elife* 2015;4:e07485.

Bonet, J.*, et al.* ArchDB 2014: structural classification of loops in proteins. *Nucleic Acids Res* 2014;42(Database issue):D315-319.

Erijman, A., Rosenthal, E. and Shifman, J.M. How structure defines affinity in protein-protein interactions. *PloS One* 2014;9:e110085.

Feliu, E., Aloy, P. and Oliva, B. On the analysis of protein-protein interactions via knowledge-based potentials for the prediction of protein-protein docking. *Protein Science: A Publication of the Protein Society* 2011;20:529-541.

Feliu, E. and Oliva, B. How different from random are docking predictions when ranked by scoring functions? *Proteins* 2010;78:3376-3385.

Fornes, O.*, et al.* On the Use of Knowledge-Based Potentials for the Evaluation of Models of Protein-Protein, Protein-DNA, and Protein-RNA Interactions. *Adv Protein Chem Struct Biol* 2014;94:77-120.

Garcia-Garcia, J.*, et al.* Networks of ProteinProtein Interactions: From Uncertainty to Molecular Details. *Mol Inform* 2012;31(5):342-362.

Gavin, A.-C.*, et al.* Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* 2002;415:141-147.

Gohlke, H., Kiel, C. and Case, D.A. Insights into protein-protein binding by binding free energy calculation and free energy decomposition for the Ras-Raf and Ras-RalGDS complexes. *J Mol Biol* 2003;330(4):891-913.

Gromiha, M.M., Yugandhar, K. and Jemimah, S. Protein-protein interactions: scoring schemes and binding affinity. *Curr Opin Struct Biol* 2016;44:31-38.

Gumbart, J.C., Roux, B. and Chipot, C. Efficient determination of protein-protein standard binding free energies from first principles. *J Chem Theory Comput* 2013;9(8).

Horton, N. and Lewis, M. Calculation of the free energy of association for protein complexes. *Protein Science: A Publication of the Protein Society* 1992;1:169-181.

Kastritis, P.L. and Bonvin, A.M.J.J. Are scoring functions in protein-protein docking ready to predict interactomes? Clues from a novel binding affinity benchmark. *Journal of Proteome Research* 2010;9:2216-2225.

Kastritis, P.L.*, et al.* Proteins feel more than they see: fine-tuning of binding affinity by properties of the non-interacting surface. *Journal of Molecular Biology* 2014;426:2632-2652.

Lensink, M.F.*, et al.* Prediction of homoprotein and heteroprotein complexes by protein docking and template-based modeling: A CASP-CAPRI experiment. *Proteins* 2016;84 Suppl 1:323-348.

Ma, X.H.*, et al.* A fast empirical approach to binding free energy calculations based on protein interface information. *Protein Engineering* 2002;15:677-681.

Margolin, A.A.*, et al.* ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics* 2006;7 Suppl 1:S7.

Marillet, S., Boudinot, P. and Cazals, F. High-resolution crystal structures leverage protein binding affinity predictions. *Proteins* 2016;84:9-20.

McCammon, J.A. Theory of biomolecular recognition. *Curr Opin Struct Biol* 1998;8(2):245-249.

Méndez, R.*, et al.* Assessment of blind predictions of protein-protein interactions: current status of docking methods. *Proteins* 2003;52:51-67.

Moal, I.H., Agius, R. and Bates, P.A. Protein-protein binding affinity prediction on a diverse set of structures. *Bioinformatics (Oxford, England)* 2011;27:3002-3009.

Moal, I.H. and Bates, P.A. Kinetic rate constant prediction supports the conformational selection mechanism of protein binding. *PLoS computational biology* 2012;8:e1002351.

Moal, I.H., Jiménez-García, B. and Fernández-Recio, J. CCharPPI web server: computational characterization of protein-protein interactions from structure. *Bioinformatics (Oxford, England)* 2015;31:123-125.

Moritsugu, K., Terada, T. and Kidera, A. Energy landscape of all-atom protein-protein interactions revealed by multiscale enhanced sampling. *PLoS Comput Biol* 2014;10(10):e1003901.

Pedregosa, F.*, et al.* Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 2011;12:2825-2830.

Planas-Iglesias, J.*, et al.* Understanding protein-protein interactions using local structural features. *Journal of Molecular Biology* 2013;425:1210-1224.

Planas-Iglesias, J.*, et al.* Cardiolipin Interactions with Proteins. *Biophys J* 2015;109(6):1282-1294.

Planas-Iglesias, J.*, et al.* iLoops: a protein-protein interaction prediction server based on structural features. *Bioinformatics (Oxford, England)* 2013;29:2360-2362.

Plattner, N.*, et al.* Complete protein-protein association kinetics in atomic detail revealed by molecular dynamics simulations and Markov modelling. *Nature Chemistry* 2017;doi:10.1038/nchem.2785.

Robinson, C.V., Sali, A. and Baumeister, W. The molecular sociology of the cell. *Nature* 2007;450:973-982.

Rodriguez, R.A., Yu, L. and Chen, L.Y. Computing Protein-Protein Association Affinity with Hybrid Steered Molecular Dynamics. *J Chem Theory Comput* 2015;11(9):4427-4438.

Schlosshauer, M. and Baker, D. Realistic protein-protein association rates from a simple diffusional model neglecting long-range interactions, free energy barriers, and landscape ruggedness. *Protein Science: A Publication of the Protein Society* 2004;13:1660-1669.

Schneidman-Duhovny, D.*, et al.* PatchDock and SymmDock: servers for rigid and symmetric docking. *Nucleic Acids Research* 2005;33:W363-367.

Schreiber, G., Haran, G. and Zhou, H.-X. Fundamental aspects of protein-protein association kinetics. *Chemical Reviews* 2009;109:839-860.

Segura, J.*, et al.* VORFFIP-driven dock: V-D2OCK, a fast and accurate protein docking strategy. *PLoS One* 2015;10(3):e0118107.

Selzer, T., Albeck, S. and Schreiber, G. Rational design of faster associating and tighter binding protein complexes. *Nature Structural Biology* 2000;7:537-541.

Su, Y.*, et al.* Quantitative prediction of protein-protein binding affinity with a potential of mean force considering volume correction. *Protein Science: A Publication of the Protein Society* 2009;18:2550-2558.

Tang, C., Iwahara, J. and Clore, G.M. Visualization of transient encounter complexes in protein-protein association. *Nature* 2006;444:383-386.

Tian, F., Lv, Y. and Yang, L. Structure-based prediction of protein-protein binding affinity with consideration of allosteric effect. *Amino Acids* 2012;43:531-543.

Ubbink, M. The courtship of proteins: understanding the encounter complex. *FEBS letters* 2009;583:1060-1066.

Vangone, A. and Bonvin, A.M.J.J. Contacts-based prediction of binding affinity in protein-protein complexes. *eLife* 2015;4:e07454.

Vreven, T.*, et al.* Prediction of protein-protein binding free energies. *Protein Science: A Publication of the Protein Society* 2012;21:396-404.

Vreven, T., *et al.* Updates to the integrated protein-protein interaction benchmarks: Docking benchmark version 5 and affinity benchmark version 2. *Journal of molecular biology* 2015;427:3031-3041.

Wass, M.N., *et al.* Towards the prediction of protein interaction partners using physical docking. *Molecular Systems Biology* 2011;7:469.

Zhou, H.-X. and Bates, P.A. Modeling protein association mechanisms and kinetics. *Current Opinion in Structural Biology* 2013;23:887-893.

Zhou, H.X. Brownian dynamics study of the influences of electrostatic interaction and diffusion on protein-protein association kinetics. *Biophysical Journal* 1993;64:1711-1726.